An Analysis of Credits to Graduation at the University of Utah

Jeremy Morris

morris@math.utah.edu

University of Utah

Outline

1. Purpose
2. Data collection and cleaning
3. Credits at graduation
4. Testing the exponential assumption
5. Generalized likelihood ratio test for homogeneity of the mean
6. Regression obeying two different regimes
7. Conclusions

Purpose

- Statistical investigation of credits to graduation.
- Inform Faculty, Staff and Administrators of how credits to graduation has been changing over time.
- Make some predictions about visible trends.

Data Collection and Cleaning

- Data collected each year for all graduating students from July 1 of previous year to June 30 of current year.
- We are interested only in first time undergraduates meeting graduation requirements.
- Graduation requirements specify 122 semester credits for graduation. This requirement was relaxed to 120 semester credits on advice from the administration.
- The U went from a quarter system to a semester system in 1999, all quarter credits were adjusted for this change.
- The minimum requirement of 120 credits was removed so that we are considering excess credits at the time of graduation.

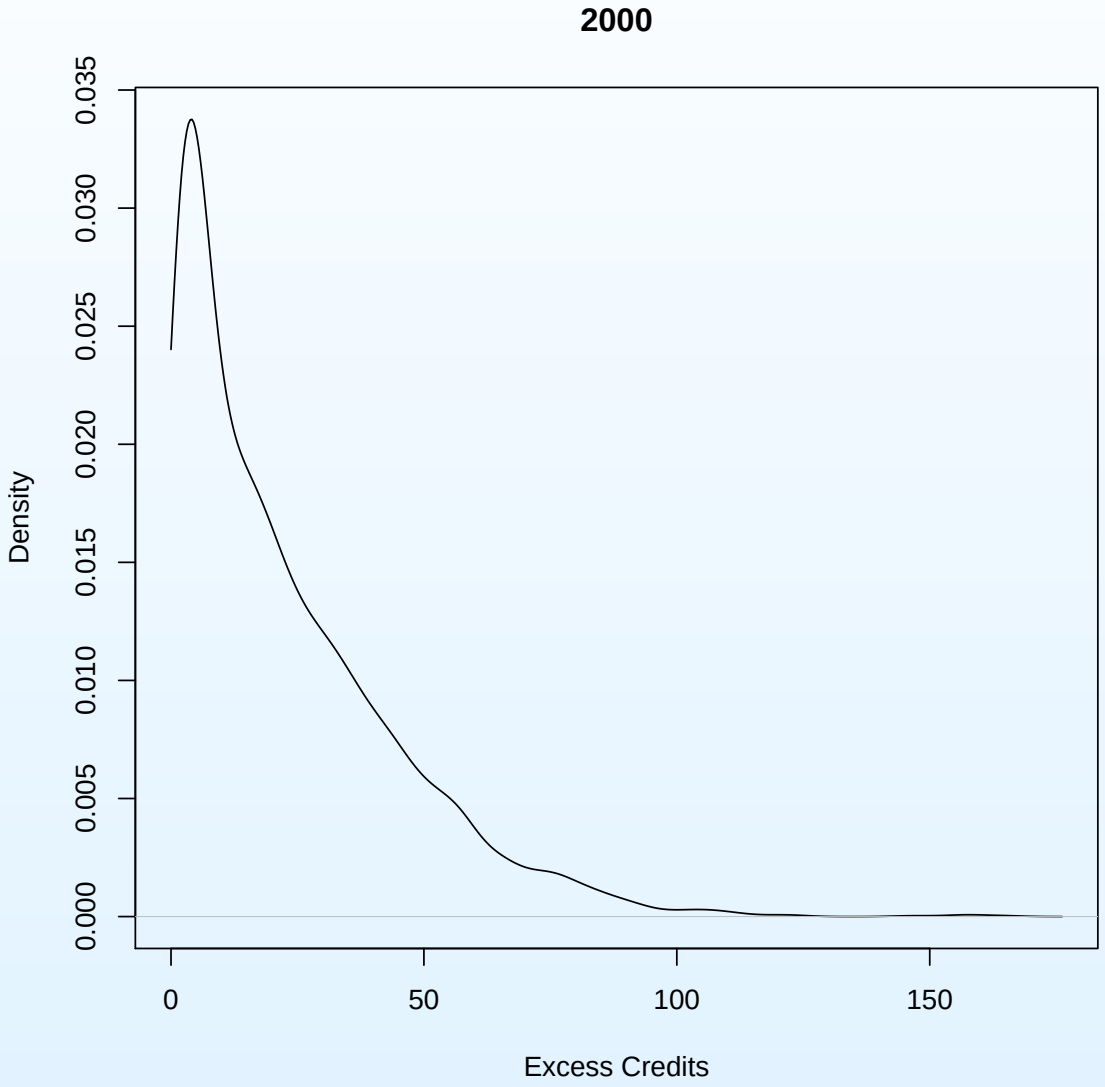
Distribution Assumptions for Credits to Graduation

- Credits to graduation are initially assumed to be exponentially distributed.
- This assumption is made because students continue to take courses until the requirements for their program are met. This represents a waiting time.
- We will look at the mean and standard deviation which are supposed to be the same under this assumption.

Table of Mean and Standard Deviations

Year	Mean	St Deviation
1997	22.119	21.436
1998	21.828	22.066
1999	22.465	22.595
2000	22.359	21.648
2001	23.009	21.983
2002	22.985	22.671
2003	23.533	23.007
2004	24.009	23.392
2005	24.544	23.886
2006	25.148	24.670

Density Plot for 2006



The χ^2 Test

- Basic idea is to cut the range of the data into evenly spaced cells and count how many observations fall in each cell. Let

$$Y_i = \sum_{j=1}^{n} I\{t_{i-1} < X_j \leq t_i\}, \qquad 1 \leq i \leq K$$

- Then we have the test statistic

$$Q = \sum_{i=1}^{K} \frac{(Y_i - n[F_0(t_i, \hat{\theta}) - F_0(t_{i-1}, \hat{\theta})])^2}{n[F_0(t_i, \hat{\theta}) - F_0(t_{i-1}, \hat{\theta})]}$$

- Then $Q \sim \chi^2(K - d - 1)$. Where $d = 1$ because we have the maximum likelihood estimate $\hat{\theta} = \bar{X}$.
- We reject for large values of Q.

Results from the χ^2 Test

Year	Q	df	p value
1997	1.721	16	1.000
1998	2.838	14	0.999
1999	3.770	13	0.993
2000	2.015	15	1.000
2001	0.961	8	0.998
2002	0.925	9	1.000
2003	2.762	17	1.000
2004	0.885	11	1.000
2005	1.097	8	0.998
2006	0.995	8	0.998
All	5.462	11	0.907

Transformation Into Uniform Order Statistics

- Let $S(n) = X_1 + X_2 + \cdots + X_n$. Then we have

$$\left(\frac{S(1)}{S(n)}, \frac{S(2)}{S(n)}, \ldots, \frac{S(n-1)}{S(n)} \right)$$

- General results on the uniform empirical distribution give the Cramér-von Mises statistic

$$CM = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0(x_i; \hat{\theta}) - \frac{i - 0.5}{n} \right)^2$$

  and the Kolmogorov-Smirnov statistic

$$D = \max_{1 \leq i \leq n} |F_0(x_i) - S_n(x_i)|$$

- We reject for large values of both.

Results for Uniform Order Statistics Tests

Year	D	Critical Value	CM	Critical Value
1997	0.0787	0.0564	1.56	0.224
1998	0.0779	0.0556	1.29	0.224
1999	0.0646	0.0314	2.13	0.224
2000	0.0734	0.0294	2.44	0.224
2001	0.0755	0.0287	1.90	0.224
2002	0.0855	0.0275	2.11	0.224
2003	0.0885	0.0276	2.37	0.224
2004	0.0798	0.0294	1.88	0.224
2005	0.0796	0.0309	2.39	0.224
2006	0.0938	0.0355	2.00	0.224
All	0.0787	0.0150	17.38	0.224

Total Time on Test Transformation

- The Total Time on Test Transformation is given as

$$T_k = \frac{\sum_{i=1}^{k}(n-i+1)(X_{i+1,n} - X_{i,n})}{\sum_{i=1}^{n-1}(n-i+1)(X_{i+1,n} - X_{i,n})}, \quad 1 \leq k \leq n-1.$$

- Then we have the test statistics

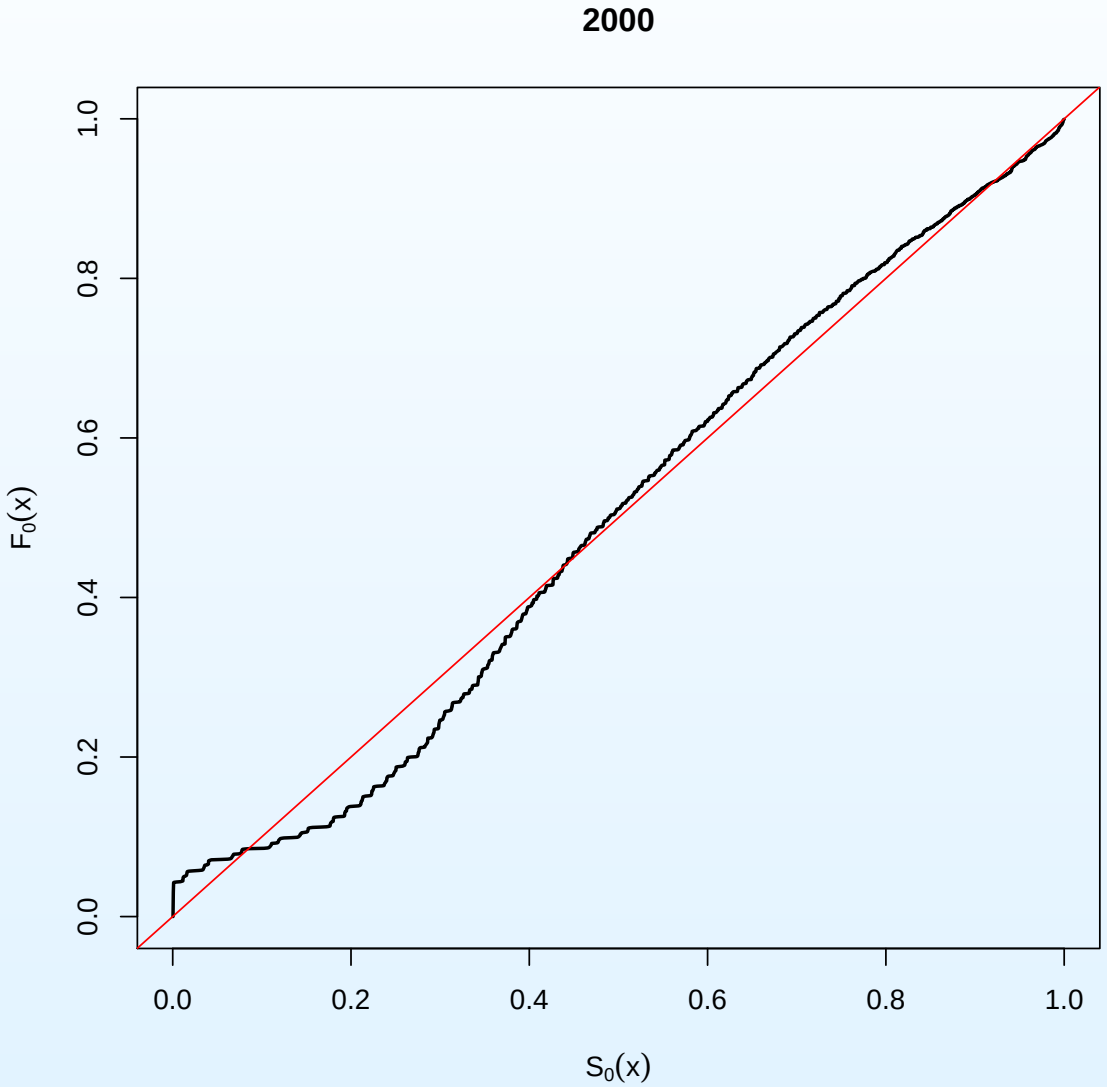
$$t_1 = \sqrt{n} \max_{1 \leq k \leq n-1} \left| T_k - \frac{k}{n} \right| \quad \xrightarrow{d} \quad \sup_{0 \leq t \leq 1} |B(t)|$$

$$t_2 = \sum_{k=1}^{n-1} \left(T_k - \frac{k}{n} \right)^2 \quad \xrightarrow{d} \quad \int_0^1 B^2(t) \quad dt,$$

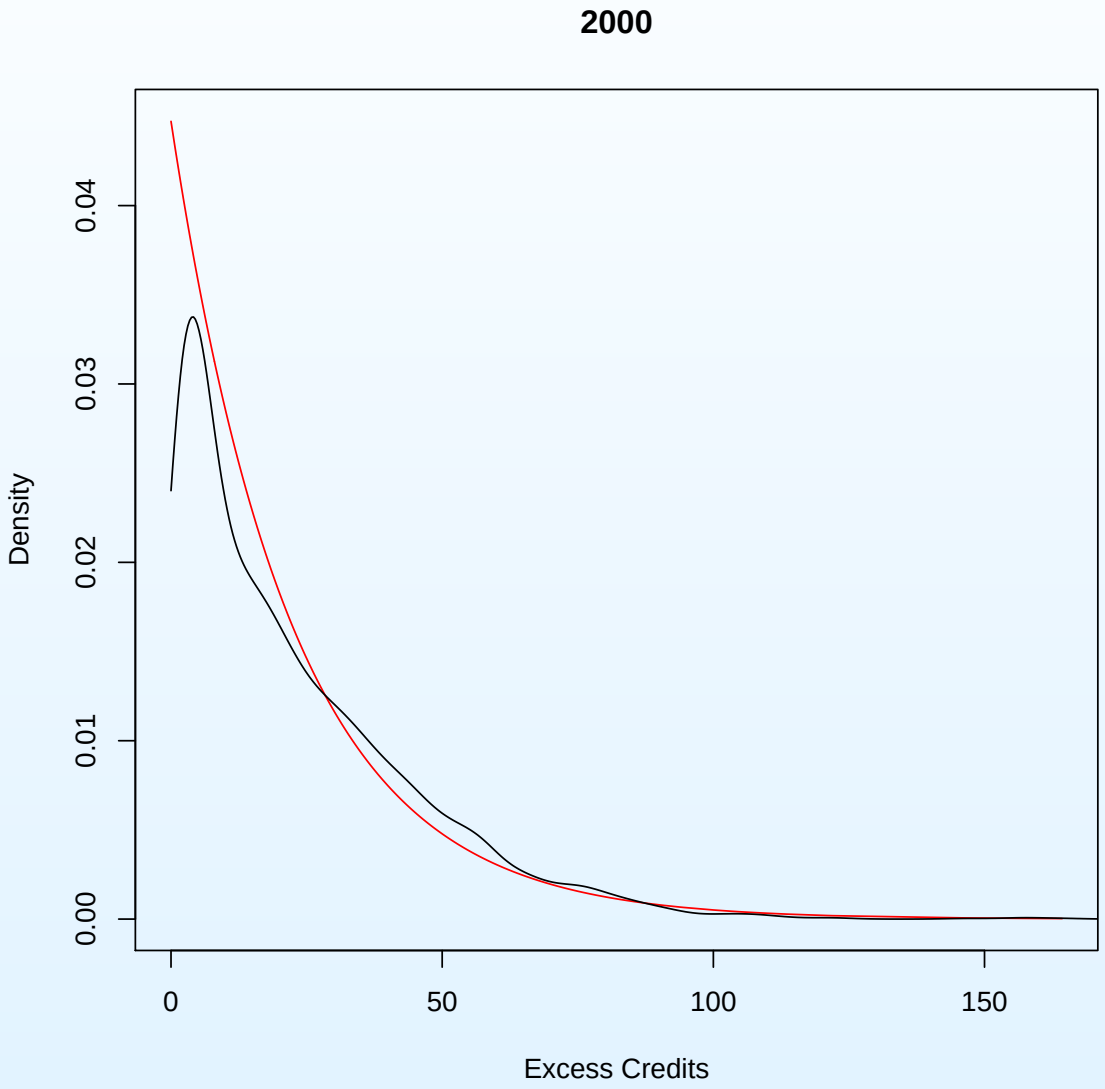
Results for Total Time on Test Transformation Tests

Year	t_1	t_2
1997	2.906	3.946
1998	2.805	4.038
1999	3.709	5.855
2000	4.016	7.547
2001	3.534	6.060
2002	3.559	4.905
2003	3.669	6.217
2004	3.617	5.382
2005	4.008	6.714
2006	3.930	6.154
All	9.835	50.233

Why do we reject the exponential assumption?



Why do we reject the exponential assumption? (cont.)



Gamma Distribution

- The gamma distribution has the density function

$$f(x) = \frac{1}{\theta^{\kappa}\Gamma(\kappa)} x^{\kappa-1} e^{-x/\theta}, \quad x > 0$$

- The gamma distribution is an abstraction of the exponential distribution and represents the sum of independent exponential random variables.
- Gamma($1, \theta$) = Exp(θ)

The Parameters of the Gamma Distribution

- Maximum likelihood estimator for θ is given as

$$\hat{\theta} = \frac{\bar{x}}{\hat{\kappa}}$$

- For κ the maximum likelihood estimate is the solution to

$$\log \hat{\kappa} - \Psi(\hat{\kappa}) - \log \bar{x}/\tilde{x} = 0$$

where $\tilde{x}$ is the geometric mean of the data and

$$\Psi(x) = \frac{\Gamma'(\kappa)}{\Gamma(\kappa)}$$

Parameters continued

- We use

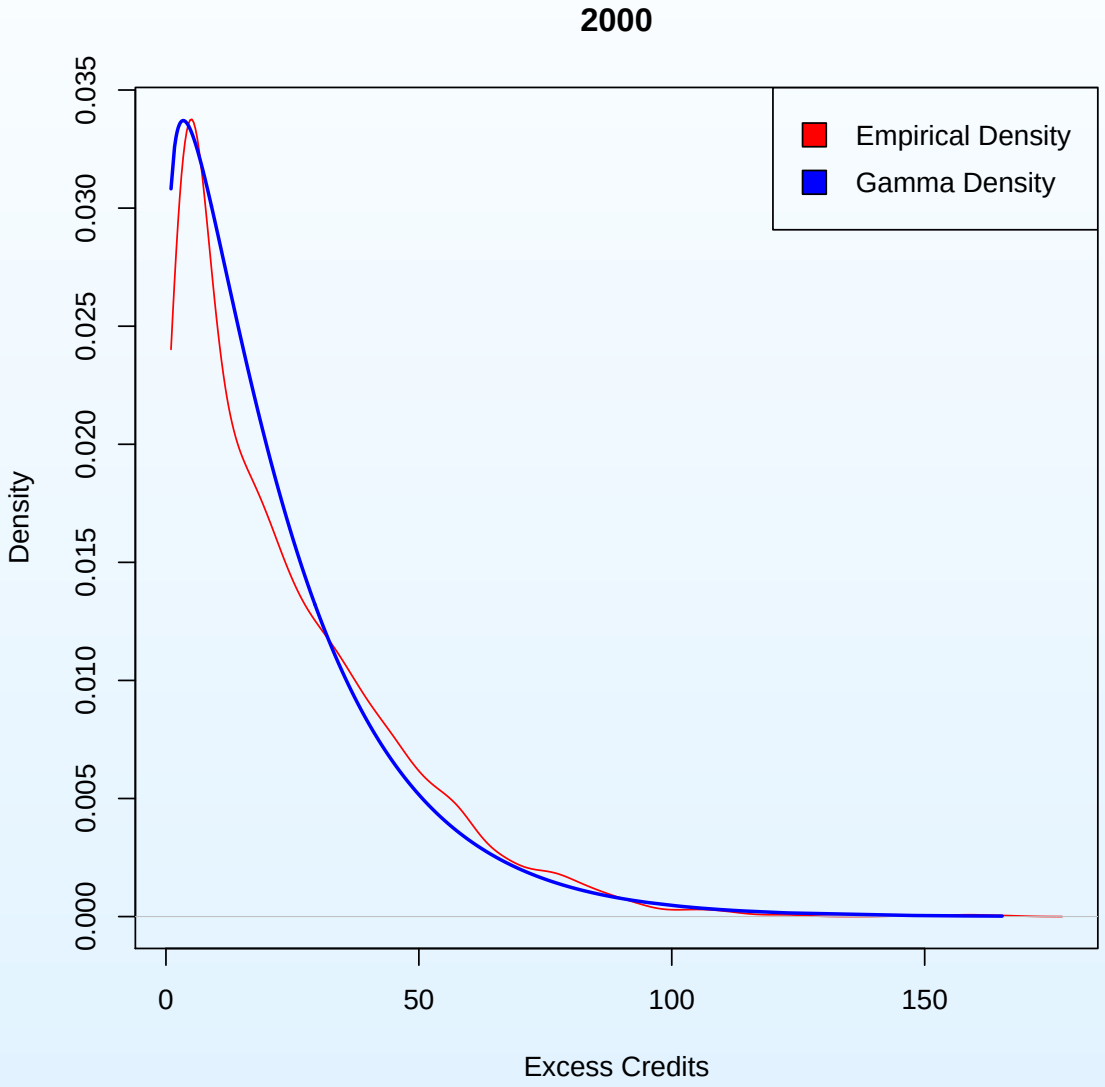
$$\hat{\kappa} = \frac{0.5000876 + 0.1648852M - 0.0544274M^2}{M}, \quad 0 < M \leq 0.5772$$

where $M = \log \bar{x}/\tilde{x}$

Parameter Values

Year	M	$\hat{\kappa}$	$\hat{\theta}$
1997	0.46180	1.2227	18.909
1998	0.47887	1.1831	19.294
1999	0.49017	1.1584	20.256
2000	0.48407	1.1716	19.938
2001	0.48335	1.1732	20.464
2002	0.50057	1.1367	21.101
2003	0.49979	1.1383	21.553
2004	0.48835	1.1623	21.516
2005	0.49477	1.1487	22.237
2006	0.50994	1.1178	23.392
All	0.49107	1.1565	21.013

Empiricial Density vs. Gamma Density



Testing the Mean

- We make the following assumption

$$X_{ij} \sim Exp(\theta_i) \quad 1 \leq i \leq k, \quad 1 \leq j \leq n_i$$

- Then we want to test the hypothesis

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k \qquad H_a : \theta_1 \neq \theta_2 \neq \ldots \neq \theta_k$$

- We use a generalized likelihood ratio test, defined as

$$\Lambda(\boldsymbol{X}) = \frac{\max\limits_{\theta \in \Omega_0} f(\boldsymbol{X}; \theta)}{\max\limits_{\theta \in \Omega} f(\boldsymbol{X}; \theta_1, \ldots, \theta_k)} = \frac{f(\boldsymbol{X}; \hat{\theta})}{f(\boldsymbol{X}; \hat{\theta}_1, \ldots, \hat{\theta}_k)}$$

Log-likelihood Function

- The log-likelihood is given as

$$\log \Lambda(\boldsymbol{X}) = \log f(\boldsymbol{X}; \hat{\theta}) - \log f(\boldsymbol{X}; \hat{\theta}_1, \ldots, \hat{\theta}_k)$$

- If H_0 holds, then $-2 \log \Lambda(\boldsymbol{X}) \sim \chi^2_{k-1}$
- We reject for large values of $-2 \log \Lambda(\boldsymbol{X})$.
- The maximum likelihood estimates for this test are

$$\hat{\theta} = \left(\sum_{k=1}^{k} n_i \right)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} \quad \text{and}$$

$$\hat{\theta}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

Log-liklihood and Test Statistic

- Then the log-likelihood function has the form

$$\log \Lambda(\boldsymbol{X}) = \sum_{i=1}^{k} n_i \log \hat{\theta}_i - \left(\sum_{i=1}^{k} n_i \right) \log \hat{\theta}$$

- In practice we use

$$T = \sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i - \hat{\theta}}{\hat{\theta}} \right)^2$$

Results

- We test all of the data together and get a p value of 1.084847×10^{-10}. We reject the hypothesis that the mean stays the same for all ten years.
- We would like to know when the mean changed. To do this, we test the data for the first two years, if the test does not reject, we add another until the test rejects.
- We use the Bonferroni method to determine what the significance level should be.

The Bonferonni Method

- Let C_i denote the i^{th} test performed ($i = 1, 2, \ldots, m$). We would like to have joint coverage for all m tests so that

$$P\{C_i \text{ true}\} = 1 - \alpha_i, \quad i = 1, 2, \ldots, m.$$

  By application of the Bonferonni inequality we get

$$P\{\text{all } C_i \text{ true}\} \geq 1 - \sum_{i=1}^{m} \alpha_i$$

  where $\alpha = \sum_{i=1}^{m} \alpha_i$.

- If we would like $\alpha = 0.05$ and expect that there will be two intervals, we need to reject (or fail to reject) both tests at $\alpha_i = 0.025$.

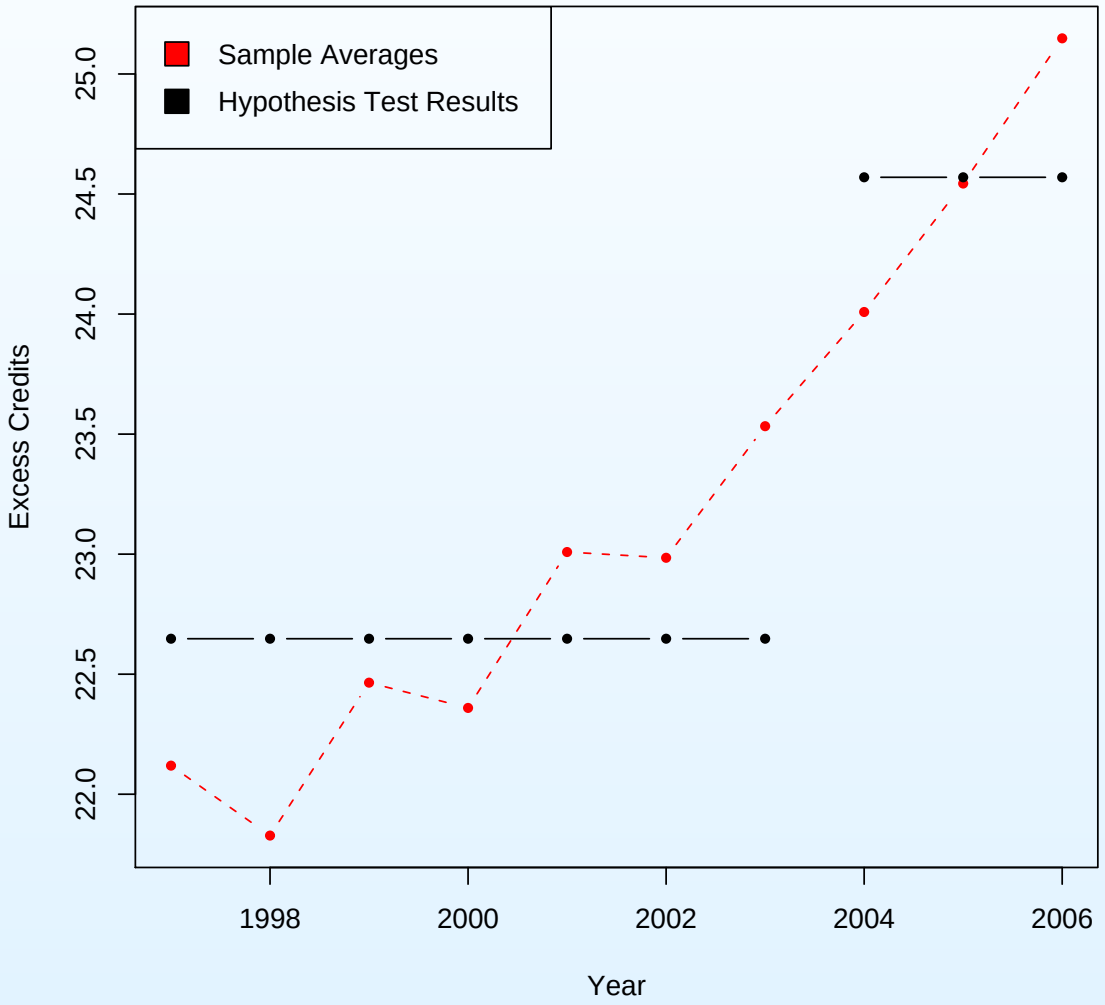
Sequential Test Results

- The p-values for these tests are

Start Year	End Year	p-value
1997	2004	0.001
1997	2003	0.047
2004	2006	0.148

- We reject that the mean remains the same from 1997 through 2004.
- We fail to reject that the mean remains the same from 1997 through 2003 and again from 2004 to 2006.

Graphic Results of Sequential Test



Regression obeying two different regimes

- We would like to test the hypothesis (H_0) that there is only one regression equation, or

$$y_i = \alpha x_i + \beta + \varepsilon_i$$

- Against the alternative hypothesis (H_A) that there are two regression equations, or

$$y_i = \alpha_1 x_i + \beta_1 + \varepsilon_i \qquad 1 \leq i \leq k^*$$
$$y_i = \alpha_2 x_i + \beta_2 + \varepsilon_i \qquad k^* < i \leq N$$

Generalized Likelihood Ratio Tests

- We assume the average excess credits are normally distributed for each year.
- We will derive tests under the following scenarios
  1. Variances unequal and unknown.
  2. Variances equal and unknown.
  3. Variances equal and known.
- We use the likelihood ratio

$$\Lambda_k = \max_{2 \leq k \leq N-2} \frac{L_k(\boldsymbol{y})}{L(\boldsymbol{y})}$$

and reject H_0 for large values of Λ_k.

Variances unequal and unknown

- Log-likelihood equation under H_A

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{k}{2}\log \hat{\sigma}^2_{1,k} - \frac{N-k}{2}\log \hat{\sigma}^2_{2,k} - \frac{N}{2}.$$

- Under H_0, we have

$$\ell(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \hat{\sigma}^2 - \frac{N}{2}$$

- And finally, we have

$$\lambda_k = \frac{N}{2}\log \hat{\sigma}^2 - \frac{k}{2}\log \hat{\sigma}^2_{1,k} - \frac{N-k}{2}\log \hat{\sigma}^2_{2,k}$$

Variances equal and unknown

- Log-likelihood equation under H_A

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \hat{\sigma}_k^2 - \frac{N}{2}$$

- Under H_0, we have

$$\ell(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \hat{\sigma}^2 - \frac{N}{2}$$

- And finally, we have

$$\lambda_k(\boldsymbol{y}) = \frac{N}{2}\log \hat{\sigma}^2 - \frac{N}{2}\log \hat{\sigma}_k^2.$$

Variances equal and known

- Log-likelihood equation under H_A

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i - \hat{\beta}_{1,k})^2$$

$$- \frac{1}{2\sigma^2}\sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{2,k}x_i - \hat{\beta}_{2,k})^2$$

- Under H_0, we have

$$\ell(\boldsymbol{y}; \hat{\alpha}, \hat{\beta}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\alpha}x_i - \hat{\beta})^2,$$

Variances equal and known (continued)

- Finally, the log-likelihood equation is

$$\lambda_k = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\alpha}x_i - \hat{\beta})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i - \hat{\beta}_{1,k})^2$$
$$- \frac{1}{2\sigma^2}\sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{2,k}x_i - \hat{\beta}_{2,k})^2$$

- In this scenario we use $\sigma^2 = 0.04$. This value was taken from the variance of the original samples divided by n_i, or the number of students in each year.

Critical Values

- In each of the scenarios we reject if

$$T_k = \max_{2 \leq k \leq 9} \lambda_k$$

  is large.
- We know that 2λ would be χ^2 if the change point k were known.
- k is not known, so we will use a resampling technique to estimate the critical values for this test.

Critical Values (continued)

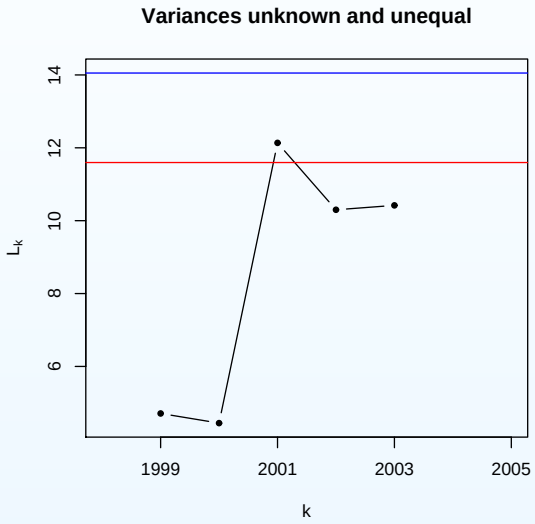
- To find the critical values, we use

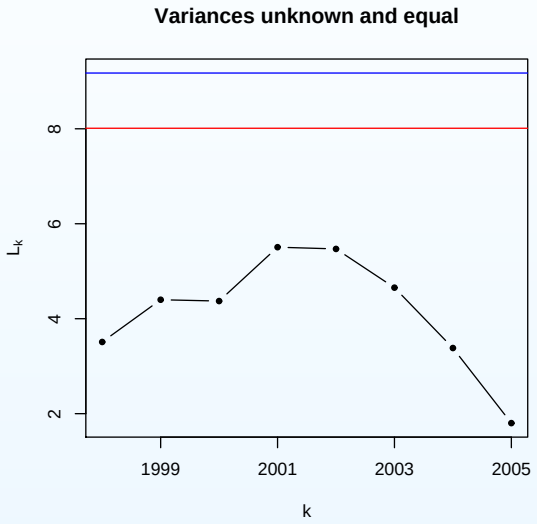
$$y_i = \beta + \alpha x_i + \varepsilon_i, \quad 1 \leq i \leq 10,$$

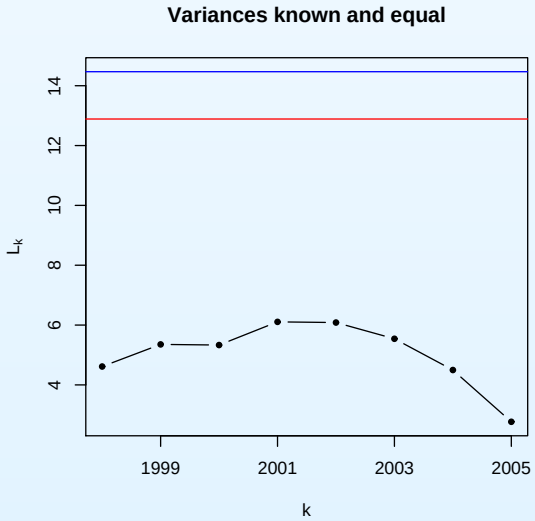
  where the ε_i will be independent normally distributed random numbers with mean 0 and variance $\sigma^2 = 0.04$.
- Then we calculate $T_{k,n}$ $(1 \leq n \leq 1000)$. And take the $1 - \alpha$ percentile as the critical values.

	$\alpha = 0.10$	$\alpha = 0.05$
Variance Unequal and Unknown	11.596	14.052
Variance Equal and Unknown	8.011	9.174
Variance Equal and Known	12.886	14.468

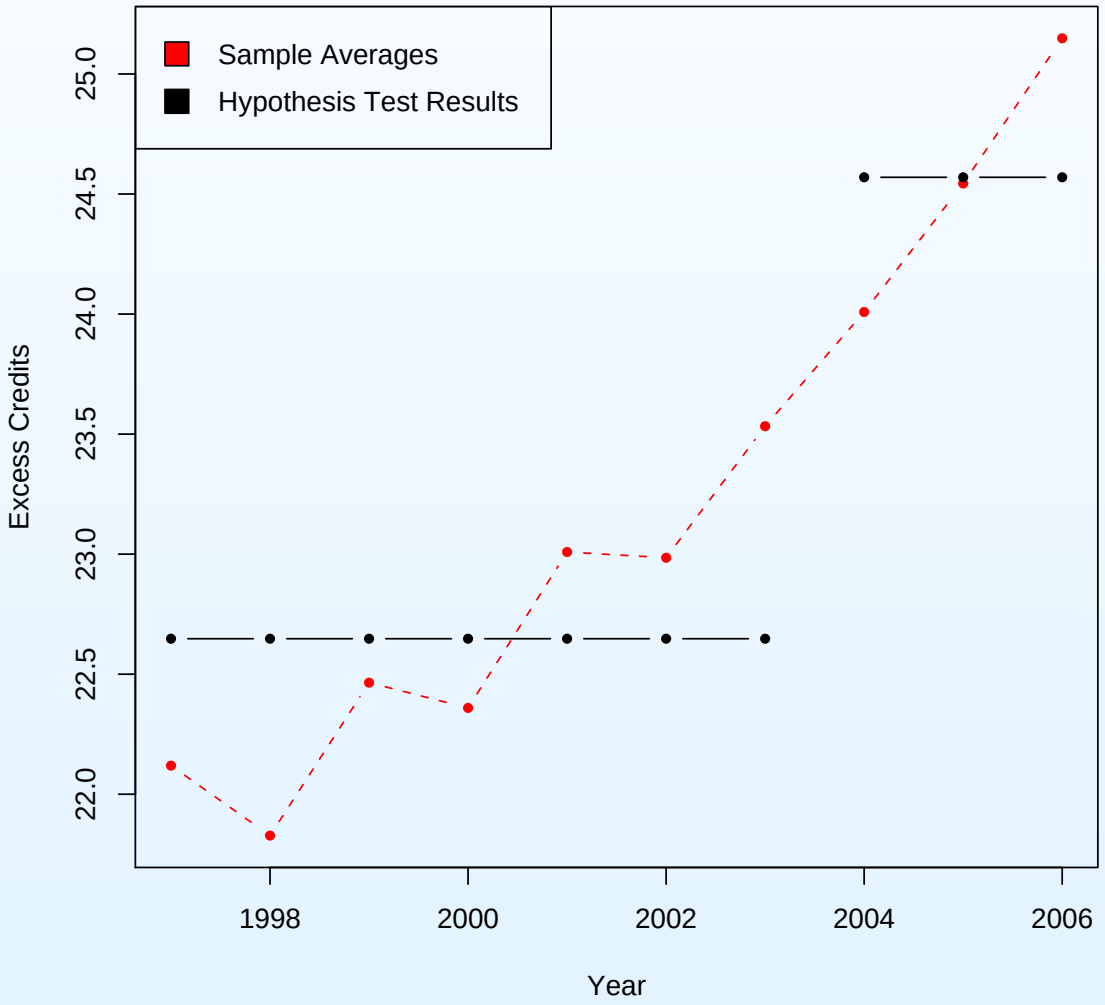
Regime Change Results







Graphic Results of Sequential Test



Verification of Results

- What if we think that there really was a regime shift in 2001?
- Consider two models. In each case we write the models in the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{Y}' = (22.12, 21.83, 22.47, 22.36, 23.01, 22.99, 23.53, 24.01, 24.54, 25.15)$$

Model 1

- In this model we assume one regression line.
- The design matrix is given as

$$\boldsymbol{X} = \begin{pmatrix} 1997 & 1 \\ 1998 & 1 \\ 1999 & 1 \\ 2000 & 1 \\ 2001 & 1 \\ 2002 & 1 \\ 2003 & 1 \\ 2004 & 1 \\ 2005 & 1 \\ 2006 & 1 \end{pmatrix}$$

Model 2

- In the second model we assume two regression lines with a split between 2001 and 2002.
- The design matrix is given as

$$\boldsymbol{X} = \begin{pmatrix} 1997 & 1 & 0 & 0 \\ 1998 & 1 & 0 & 0 \\ 1999 & 1 & 0 & 0 \\ 2000 & 1 & 0 & 0 \\ 2001 & 1 & 0 & 0 \\ 0 & 0 & 2002 & 1 \\ 0 & 0 & 2003 & 1 \\ 0 & 0 & 2004 & 1 \\ 0 & 0 & 2005 & 1 \\ 0 & 0 & 2006 & 1 \end{pmatrix}.$$

Results

- For Model 1 we get

$$\boldsymbol{\beta}' = (0.348, -674.207)$$

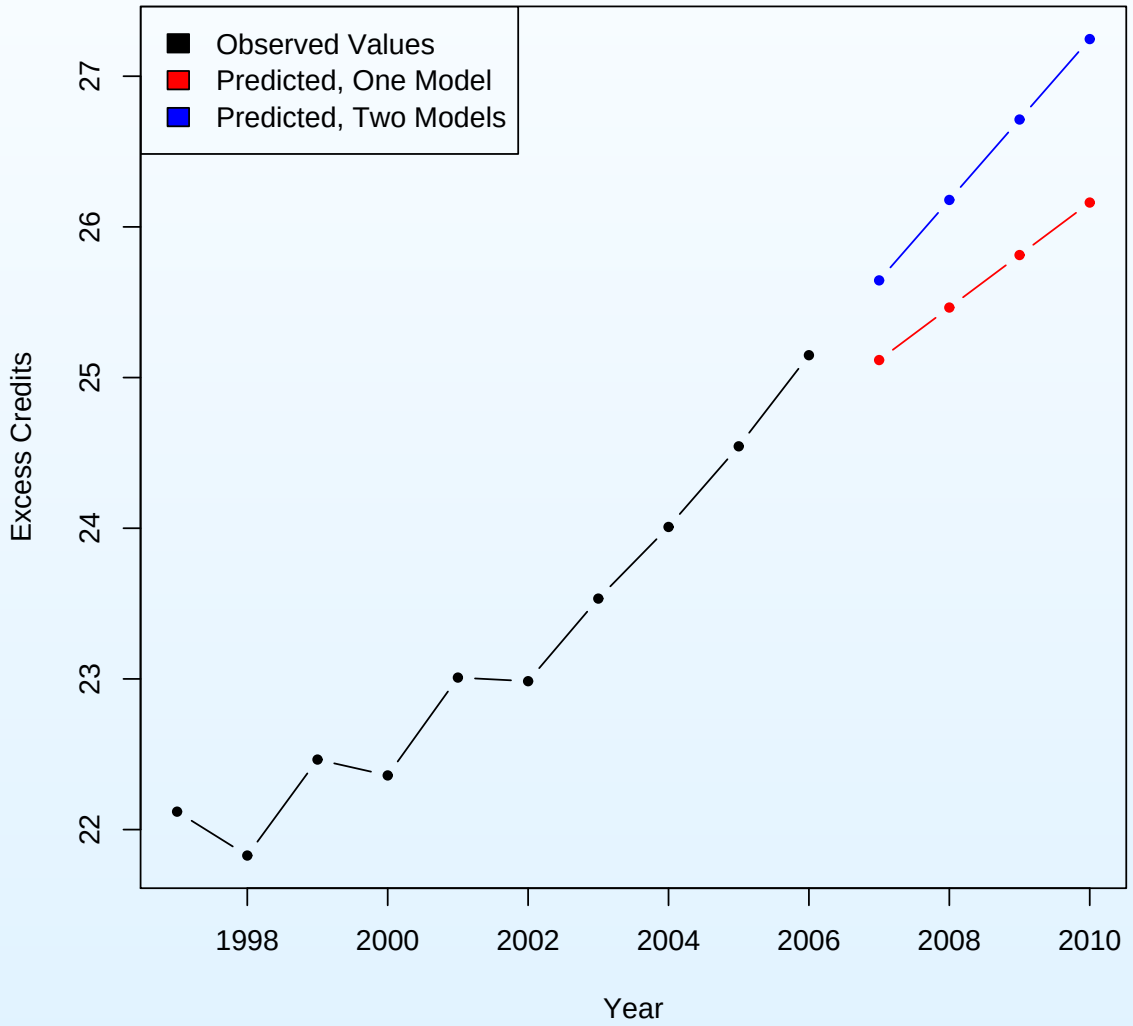
  with $R^2 = 0.9319$.

- For Model 2 we get

$$\boldsymbol{\beta}' = (0.231, -439.683, 0.533, -1045.667)$$

  with $R^2 = 1$.

Projections



Conclusions and Suggestions for Further Research

- Conclusions
  1. Students credit hours are increasing.
  2. At worst, students are taking one extra course every six years.
- Further research could be done to try and determine the cause of the increase. Some suggestions are
  1. Economics
  2. Work status of students
  3. Changing college/department requirements
  4. Student movement between colleges/departments.
  5. Number of certificates, double majors and combined BS/MS programs

Compare Avg Credits by College